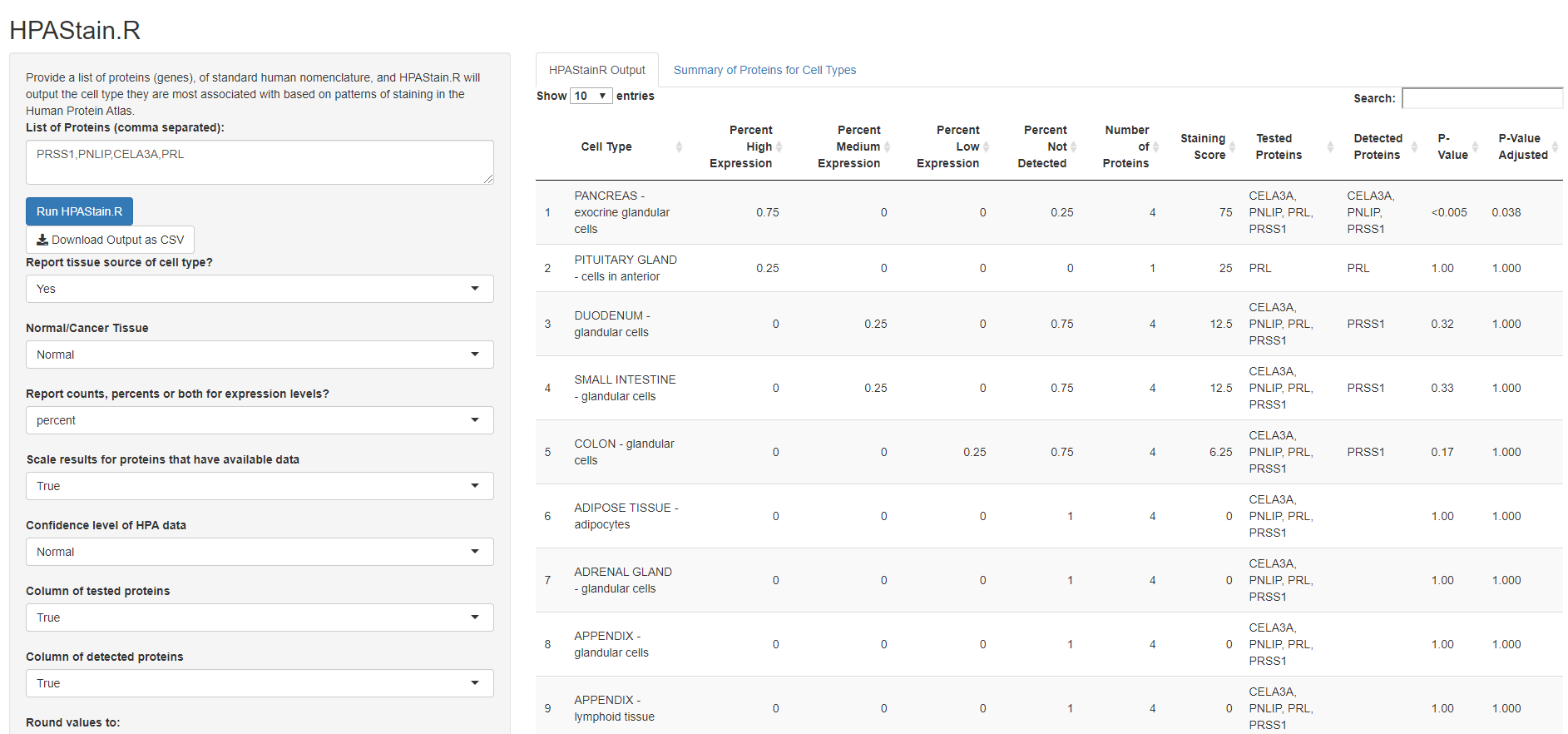
**Introduction/Abstract**

The Human Protein Atlas (HPA) is a resource that has accumulated deep levels of protein staining in various human tissues and cell types [25613900, 28818916, 32139519]. While it is easy to query the website for a specific protein of interest, there is no option to query a list of proteins to see if said list of proteins stain a specific tissue or cell type. Other gene list enrichment tools such as Enrichr have yet to incorporate this dataset into their tools as well. Introducing HPAStainR (<https://32tim32.shinyapps.io/HPAStainR/>, Figure 1), a shiny app and R function developed to query the staining data available from the Human Protein Atlas database. HPAStainR allows you to input a list of proteins/genes and returns a rank ordered list of cell types within tissues that are stained for the proteins you’ve input. There are multiple customizations allowed, such as the ability to include cancer tissue data, change the HPA confidence levels, a column telling what exact proteins from the list were detected, a p-value for how many cell type specific proteins in the list are in a cell type, and the ability to download the output as a comma separated file.

HPAStainR works by using the publically available HPA cell type scored staining data to subset out and summarize the queried proteins. It ranks on a 0 to 100 “staining score”, based on the pathologist annotated staining intensity of the proteins in the given cell types. For example, a query of the pancreatic enzymes PRSS1, PNLIP, and CELA3A, along with the protein PRL, would identify “pancreas exocrine glandular cells” as the top hit with a staining score of 75 due to the high staining intensity in 3 of the 4 genes. The second hit would be the pituitary gland due to PRL’s high expression, followed by intestinal glandular cells which only have medium staining of PRSS1. The resulting table on the shiny app can be sorted on any column to more readily answer one’s scientific question, may that be the number of proteins with high expression or just the number of proteins found.



**Figure 1.** **The user interface of the HPAStainR shiny app.**

**Results**

To show the functionality of the shinyapp we applied HPAStainR to the Pangloa Database (PanglaoDB), a hub of single cell data, with a community-curated cell-type marker database [30951143]. We wanted to investigate how well HPAStainR would mark the cell types based on PangloaDB’s annotations. We downloaded a tsv file of PanglaoDB’s cell type gene marker data, removed all mouse specific cell type markers and filtered down to only protein coding genes. We ran these remaining 146 cell types and their 3,661 marker genes through HPAStainR. The number of marker genes per cell type in PanglaoDB are variable, ranging from 1 marker in trophoblast stem cells to 216 in interneurons, and a histogram of markers per cell type showing the distribution can be found in Supplemental Figure 1.

**Generating a confidence score for comparisons between protein lists**

HPAStainR was not designed to compare multiple gene lists between one another, and so to make valuable comparisons we generated a “confidence score”. The confidence score is calculated as follows: the number of marker proteins used from 1-50 (if more than 50 proteins are used, the protein count is capped at 50) multiplied by the staining score and a scaling factor of 0.02. The “confidence score” range is 0-100 and a score of 100 indicates a cell type where all 50 or more input proteins highly stain.

**Testing how well HPAStainR works on tissues of interest**

HPAStainR remains agnostic to where a protein/gene list comes from, so if a list comes from differential expression lung bulk RNA-sequencing, that doesn’t mean lung cell types will be the top result. By finding equivalent cell types between Panglao and HPA, as annotated by a pathologist, we attempted to see how well HPAStainR performed on the ideal cell type result. We did this to show that, while the tissue of interest might not be the top result, when you focus on said tissue, HPAStainR still performs well. Therefore we included in our PangloaDB output, both the top result of HPAStainR, and the top result in the appropriate tissue.

A subset of this analysis can be seen below in Table 1 with the full results being in Supplemental Table 1. Results were ranked by confidence score demonstrating a strong correlation of higher confidence scores with more accurate cell types being ascribed to the markers. An interesting example here are adipocytes, because while trophoblastic cells are their top HPAStainR result, the adjusted enriched protein p-value for trophoblastic cells is 1 while it is 0.042 in the adipocytes. It should also be noted that there are cell types PanglaoDB has that HPA lacks such as purkinje fiber cells.

**Table 1. A subset of 10 HPAStainR results of PangloaDB cell type marker queries**. Select tissues are when the HPAstainR output is limited to the tissue or pathologist annotated closest cell type to that of PangloaDB’s cell type.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| PangloaDB Cell Type | Confidence Score | Top HPAStainR Result | Top Result Stained Score | Select Tissue Top Result | Select Tissue Top Stained Score |
| KIDNEY proximal tubule cells | 70.5 | KIDNEY - cells in tubules | 70.5 | KIDNEY - cells in tubules | 70.5 |
| HEART MUSCLE cardiomyocytes | 66 | HEART MUSCLE - myocytes | 66 | HEART MUSCLE - myocytes | 66 |
| IMMUNE SYSTEM neutrophils | 62.5 | BONE MARROW - hematopoietic cells | 62.5 | BONE MARROW - hematopoietic cells | 62.5 |
| OLFACTORY SYSTEM olfactory epithelial cells | 49.75 | FALLOPIAN TUBE - glandular cells | 49.75 | NASOPHARYNX - respiratory epithelial cells | 36 |
| CONNECTIVE TISSUE chondrocytes | 26 | TONSIL - squamous epithelial cells | 26 | SOFT TISSUE - chondrocytes | 22 |
| CONNECTIVE TISSUE adipocytes | 25.5 | PLACENTA - trophoblastic cells | 25.5 | ADIPOSE TISSUE - adipocytes | 24.25 |
| REPRODUCTIVE granulosa cells | 14.96 | PLACENTA - trophoblastic cells | 46.75 | OVARY - follicle cells | 16.5 |
| HEART MUSCLE purkinje fiber cells | 5.775 | CAUDATE - neuronal cells | 57.75 | HEART MUSCLE - myocytes | 0 |
| BRAIN cholinergic neurons | 4.69 | SMALL INTESTINE - glandular cells | 33.5 | CEREBRAL CORTEX - neuronal cells | 24.75 |
| EMBRYO trophoblast progenitor cells | 3 | PLACENTA - trophoblastic cells | 50 | tissue not found | NA |

**HPAStainR can help determine cell type populations in bulk RNA sequencing**

To show that the use of HPAStainR is not limited to single cell analyses we decided to see if this tool could revalidate the cell type driven variation in McCall et al. [27588449]. In their paper they used bulk sequencing data to generate clusters of high variance genes and then used this data to parse out the variation in the Genotype-Tissue Expression Project’s lung tissue. From their paper they determined there were two biologically driven cell type clusters, cluster A which represented type II pneumocytes and cluster B which they ascribed the respiratory epithelium and goblet cells. HPAStainR was applied to both lists and found the top results to be lung pneumocytes and bronchus respiratory epithelial cells respectively (Supplemental Table 2, 3).

**Methods**

**HPA Data acquisition and distribution**

Human Protein Atlas normal tissue and cancer tissue data was acquired from their website (<https://www.proteinatlas.org/about/download>, last visited March 28th 2020) and all analyses took place in R (version 3.6.1) using the tidyverse package.

There are some caveats in the HPA data that should be noted. The distribution of how many proteins are tested in each of the 137 cell types varies in HPA, as a result it is important to note that not all results are equal. The number of proteins tested in cell types ranges from 1 (substantia nigra) to over 15,000 (Supplemental Figure 2, Supplemental Table 4). This variation will affect how often a protein is detected in a given cell type as well (Supplemental Figure 3). The ratio of stained to tested proteins demonstrates an enrichment at both extremes of the ratio (Supplemental Figure 4). This occurs for two reasons, there are tissues with unique cell types stained only once, and if those unique cell types fail to stain on one protein that would inflate the number of cell types that “never” stain. Secondly, in samples where few proteins were selected, proteins that they chose to test had a more likely chance of staining those cell types relative to not, as they were testing those select proteins most likely with a hypothesis of them staining.

**Tissue Enriched P-value**

While we utilized all expressed proteins in our staining score, we recognize that some proteins are more cell-type specific than others. For this analysis we generated the “enriched-protein p-value”.

To calculate the enriched-protein p-value we generated a list of cell-type enriched proteins. This was done by generating a “times stained : times tested” ratio for each protein to adjust for protein scoring frequency. This ratio was used to filter out the first quartile of proteins (proteins must stain in less than 29% of the cell types they were tested in, N = 3,190) to generate our “enriched proteins” list (Supplemental Table 5). Using this list, we ran multiple X2 tests per HPAStainR input (one per cell type). The variables being “do proteins from this enriched list of proteins stain in this cell type” and “are these proteins in the list or not”. If there is an enrichment in the list of proteins for including rare stained proteins, the result will be significant in this cell types that those proteins are found in.

**PangloaDB Test**

The data from PangloaDB was downloaded at <https://panglaodb.se/markers/PanglaoDB_markers_27_Mar_2020.tsv.gz> last updated March 27th 2020. All code for the analysis can be found on github at <https://github.com/32tim32/stainR>.

**Conclusion** As large datasets of single cell RNA sequencing analysis are curated there is more insight gained in the canonical transcriptomic profiles of genes in specific cell types. While transcriptomic profiles are beneficial, having an understanding of the cellular proteome allows us to know what proteins are actually being translated. We have also shown the tool can recapitulate bulk RNA sequencing findings making it a valuable tool to understand the cellular composition of a sample. The Human Protein Atlas is an excellent resource to observe staining patterns within cells across tissues for proteins of interest. However there is no current method to query this data with multiple proteins at once, which is a gap HPAStainR now fills.

\*

P-value enriched calculation:

While we incorporate all staining into our enriched score, we are aware that some proteins are more informative than others. Some cell types uniquely express rare proteins at high or low levels, and if you are using cell type specific proteins in your search it is important to know which HPA cell types uniquely translate said proteins.

To calculate the enriched-protein p-value we first generated a list of cell-type enriched proteins. This was done by generating a “times stained : times tested” ratio for each protein to adjust for how often a protein was tested, and then only included the first quartile of proteins (proteins must stain in less than 29% of the cell types tested in, N = 3,190) in our “enriched proteins” list (Supp. Table). This method removes ubiquitously staining proteins and prevents the amount of times a protein is tested from influencing if the protein is included in the list or not. From there we ran multiple X2 tests per HPAStainR input one per cell type. The variables being “do proteins from this enriched list of proteins stain in this cell type” and “are these proteins in the list or not”. If there is an enrichment in the list of proteins for including rare stained proteins, the result will be significant in this cell types that those proteins are found in.

Explaining cell type situation:

The data distribution in cell types varies in HPA, as result it is important to note that not all results are equal. The number of proteins tested in cell types ranges from over 15,000 to just 1 as tested in the substantia nigra (figure 1, supp table 1). This variation effects how often a protein is detected in a given cell type as well, meaning those that had fewer than 50 tested were bound to have fewer than 50 stained proteins. It is also important, when you look at the ratio of stained to tested proteins you also see and enrichment at both extremes of the ratio. This occurs for two reasons, one if there is a certain tissue stained with one protein and there are unique cell types that don’t stain to that specific protein, that increases the amount of cell types that don’t stain any proteins. Secondly, in samples where few proteins were selected, proteins that they chose to test had a more likely chance of staining those cell types relative to not as they were testing those handful of proteins most likely with a hypothesis of them staining.